HW4-STAT2131

Jihang Jiang

2023-11-05

Question1

```
library(ggplot2)
Cosme_data <- read.csv("CosmeticsSales.txt",header = TRUE,sep=" ", col.names = c("Y","X1","X2","X3"))
head(Cosme_data)
```

```
##       Y  X1  X2  X3
## 1 12.85 5.6 5.6 3.8
## 2 11.55 4.1 4.8 4.8
## 3 12.78 3.7 3.5 3.6
## 4 11.19 4.8 4.5 5.2
## 5  9.00 3.4 3.7 2.9
## 6  9.34 6.1 5.8 3.4
```

Part1:

```
model_cosme1 <- lm(Y~X1, data = Cosme_data)
summary(model_cosme1)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = Cosme_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0060 -0.7919  0.1584  1.2961  3.4824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1628     0.6712   4.712 2.69e-05 ***
## X1            1.6581     0.1641  10.104 8.23e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.892 on 42 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7016
## F-statistic: 102.1 on 1 and 42 DF,  p-value: 8.231e-13
```

Part2:

```
model_cosme2 <- lm(Y~X2,data = Cosme_data)
summary(model_cosme2)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = Cosme_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4287 -1.2874  0.2027  1.0759  3.6742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8315     0.6990   4.051 0.000215 ***
## X2            1.7926     0.1769  10.135 7.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.888 on 42 degrees of freedom
## Multiple R-squared:  0.7098, Adjusted R-squared:  0.7029
## F-statistic: 102.7 on 1 and 42 DF,  p-value: 7.507e-13
```

Part3:

```r
Full_model_cosm <- lm(Y~X1+X2+X3, data = Cosme_data)
summary(Full_model_cosm)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = Cosme_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4217 -0.9115  0.0703  1.1420  3.5479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0233     1.2029   0.851   0.4000
## X1            0.9657     0.7092   1.362   0.1809
## X2            0.6292     0.7783   0.808   0.4237
## X3            0.6760     0.3557   1.900   0.0646 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.825 on 40 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7223
## F-statistic: 38.28 on 3 and 40 DF,  p-value: 7.821e-12
```

```r
model_cosmex1x3 <- lm(Y ~ X1 +X3, data = Cosme_data)
summary(model_cosmex1x3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3, data = Cosme_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5934 -1.0162  0.1808  1.1548  3.4955
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0173     1.1978   0.849   0.4006
## X1            1.5221     0.1701   8.948 3.45e-11 ***
## X3            0.7362     0.3464   2.125   0.0396 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.818 on 41 degrees of freedom
## Multiple R-squared:  0.7374, Adjusted R-squared:  0.7246
## F-statistic: 57.58 on 2 and 41 DF,  p-value: 1.242e-12
```

```r
model_cosmex2x3 <- lm(Y~X2+X3,data = Cosme_data)
summary(model_cosmex2x3)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3, data = Cosme_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1265 -0.9973  0.0202  0.9655  3.6581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0861     1.2144   0.894   0.3764
## X2            1.6577     0.1894   8.752 6.31e-11 ***
## X3            0.6205     0.3571   1.738   0.0897 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.844 on 41 degrees of freedom
## Multiple R-squared:  0.7297, Adjusted R-squared:  0.7165
## F-statistic: 55.34 on 2 and 41 DF,  p-value: 2.255e-12
```

Comment: For the simple linear model Y ~ X1 and Y ~ X2, we can see X1 and X2 both statistically influence Y, p-value < 0.05; For the marginal t-test of X1 and X2 with controlling X3, say like Y~X1+X3 and Y~X2+X3, we can see X1 and X2 also statistically influence Y, p-value < 0.05. But for the full model Y~X1+X2+X3, both X1 and X2 do not statistically influence Y, the p-values are even far larger than 0.1. That might because there exist Multicollinearity among variables, in this case, X1 and X2 may have linear relationship.

Part4:

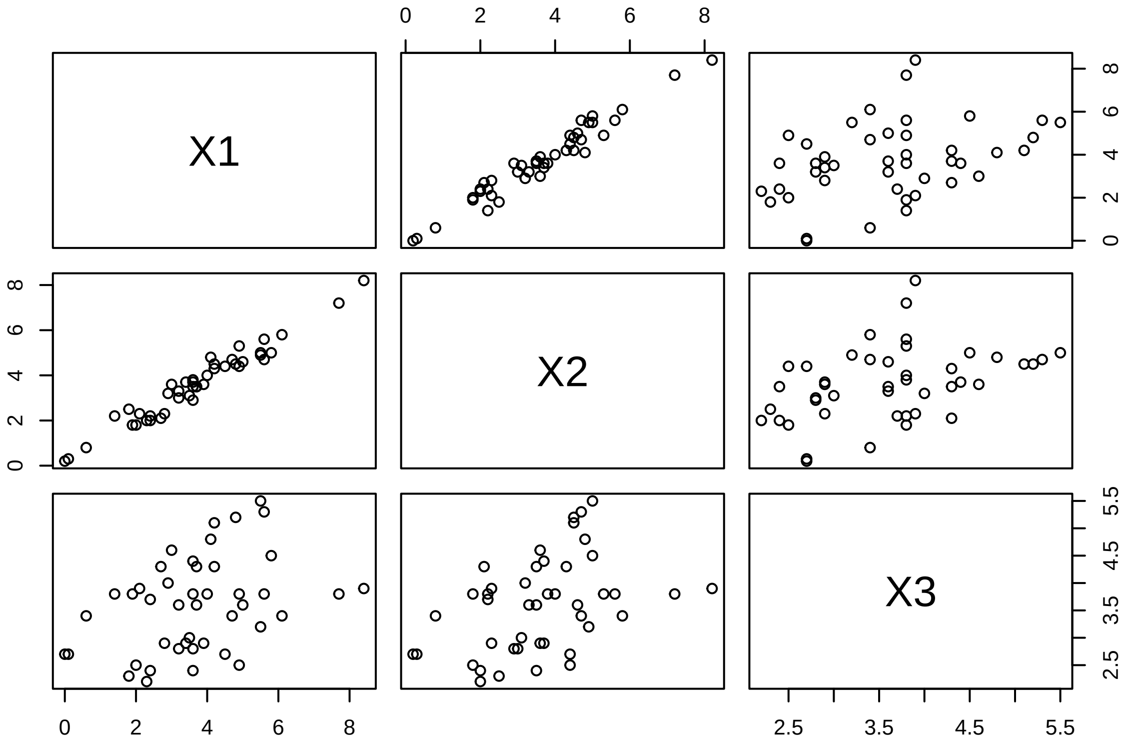
```r
library(car)
```

```
## Loading required package: carData
```

```r
vif(Full_model_cosm)
```

```
##        X1        X2        X3
## 20.072031 20.716101  1.217973
```

```r
pairs(~X1+X2+X3, data=Cosme_data)
```



We can see from the output of VIF and pairs plot, VIF values of X1 and X2 is far larger than 10, which means they have strong Multicollinearity. VIF value of X3 means X3 does not have multicollinearity with any other variables. The pair plot shows obviously X1 and X2 has linear relationship.

Problem2:

```
credit_data <- read.csv("Credit.csv",header = TRUE, sep=",")
head(credit_data)
```

```
##   X  Income Limit Rating Cards Age Education Gender Student Married Ethnicity
## 1 1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian
## 2 2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian
## 3 3 104.593  7075    514     4  71        11   Male      No      No     Asian
## 4 4 148.924  9504    681     3  36        11 Female      No      No     Asian
## 5 5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian
## 6 6  80.180  8047    569     4  77        10   Male      No      No Caucasian
##   Balance
## 1     333
## 2     903
## 3     580
## 4     964
## 5     331
## 6    1151
```

```
library(leaps)
best_subset_credit <- regsubsets(Balance~Income+Limit+Rating+Cards+Age+Education+Gender+Student+Married+
summary(best_subset_credit)
```

```
## Subset selection object
## Call: regsubsets.formula(Balance ~ Income + Limit + Rating + Cards +
##     Age + Education + Gender + Student + Married + Ethnicity,
##     data = credit_data)
```

```
## 11 Variables  (and intercept)
##                      Forced in Forced out
## Income                   FALSE      FALSE
## Limit                    FALSE      FALSE
## Rating                   FALSE      FALSE
## Cards                    FALSE      FALSE
## Age                      FALSE      FALSE
## Education                FALSE      FALSE
## GenderFemale             FALSE      FALSE
## StudentYes               FALSE      FALSE
## MarriedYes               FALSE      FALSE
## EthnicityAsian           FALSE      FALSE
## EthnicityCaucasian       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##           Income Limit Rating Cards Age Education GenderFemale StudentYes
## 1  ( 1 )  " "    " "   "*"    " "   " " " "       " "          " "       
## 2  ( 1 )  "*"    " "   "*"    " "   " " " "       " "          " "       
## 3  ( 1 )  "*"    " "   "*"    " "   " " " "       " "          "*"       
## 4  ( 1 )  "*"    "*"   " "    "*"   " " " "       " "          "*"       
## 5  ( 1 )  "*"    "*"   "*"    "*"   " " " "       " "          "*"       
## 6  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       " "          "*"       
## 7  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"          "*"       
## 8  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"          "*"       
##           MarriedYes EthnicityAsian EthnicityCaucasian
## 1  ( 1 )  " "        " "            " "               
## 2  ( 1 )  " "        " "            " "               
## 3  ( 1 )  " "        " "            " "               
## 4  ( 1 )  " "        " "            " "               
## 5  ( 1 )  " "        " "            " "               
## 6  ( 1 )  " "        " "            " "               
## 7  ( 1 )  " "        " "            " "               
## 8  ( 1 )  " "        "*"            " "               
```

```r
best_sse <- summary(best_subset_credit)
sse <- best_sse$rss
sse
```

```
## [1] 21435122 10532541  4227219  3915058  3866091  3821620  3810759  3804746
```

```r
forward_credit <- regsubsets(Balance~Income+Limit+Rating+Cards+Age+Education+Gender+Student+Married+Eth
summary(forward_credit)
```

```
## Subset selection object
## Call: regsubsets.formula(Balance ~ Income + Limit + Rating + Cards + 
##     Age + Education + Gender + Student + Married + Ethnicity, 
##     data = credit_data, nbest = 1, method = "forward")
## 11 Variables  (and intercept)
##                      Forced in Forced out
## Income                   FALSE      FALSE
## Limit                    FALSE      FALSE
## Rating                   FALSE      FALSE
## Cards                    FALSE      FALSE
## Age                      FALSE      FALSE
## Education                FALSE      FALSE
## GenderFemale             FALSE      FALSE
```

```
## StudentYes              FALSE      FALSE
## MarriedYes              FALSE      FALSE
## EthnicityAsian          FALSE      FALSE
## EthnicityCaucasian      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          Income Limit Rating Cards Age Education GenderFemale StudentYes
## 1  ( 1 ) " "    " "   "*"    " "   " " " "       " "          " "
## 2  ( 1 ) "*"    " "   "*"    " "   " " " "       " "          " "
## 3  ( 1 ) "*"    " "   "*"    " "   " " " "       " "          "*"
## 4  ( 1 ) "*"    "*"   "*"    " "   " " " "       " "          "*"
## 5  ( 1 ) "*"    "*"   "*"    "*"   " " " "       " "          "*"
## 6  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       " "          "*"
## 7  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"          "*"
## 8  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"          "*"
##          MarriedYes EthnicityAsian EthnicityCaucasian
## 1  ( 1 ) " "        " "            " "
## 2  ( 1 ) " "        " "            " "
## 3  ( 1 ) " "        " "            " "
## 4  ( 1 ) " "        " "            " "
## 5  ( 1 ) " "        " "            " "
## 6  ( 1 ) " "        " "            " "
## 7  ( 1 ) " "        " "            " "
## 8  ( 1 ) " "        "*"            " "
```

```r
forward_sse <- summary(forward_credit)$rss
forward_sse
```

```
## [1] 21435122 10532541  4227219  4032502  3866091  3821620  3810759  3804746
```

```r
backward_credit <- regsubsets(Balance~Income+Limit+Rating+Cards+Age+Education+Gender+Student+Married+Eth
summary(backward_credit)
```

```
## Subset selection object
## Call: regsubsets.formula(Balance ~ Income + Limit + Rating + Cards +
##     Age + Education + Gender + Student + Married + Ethnicity,
##     data = credit_data, nbest = 1, method = "backward")
## 11 Variables  (and intercept)
##                    Forced in Forced out
## Income              FALSE      FALSE
## Limit               FALSE      FALSE
## Rating              FALSE      FALSE
## Cards               FALSE      FALSE
## Age                 FALSE      FALSE
## Education           FALSE      FALSE
## GenderFemale        FALSE      FALSE
## StudentYes          FALSE      FALSE
## MarriedYes          FALSE      FALSE
## EthnicityAsian      FALSE      FALSE
## EthnicityCaucasian  FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          Income Limit Rating Cards Age Education GenderFemale StudentYes
## 1  ( 1 ) " "    "*"   " "    " "   " " " "       " "          " "
## 2  ( 1 ) "*"    "*"   " "    " "   " " " "       " "          " "
```

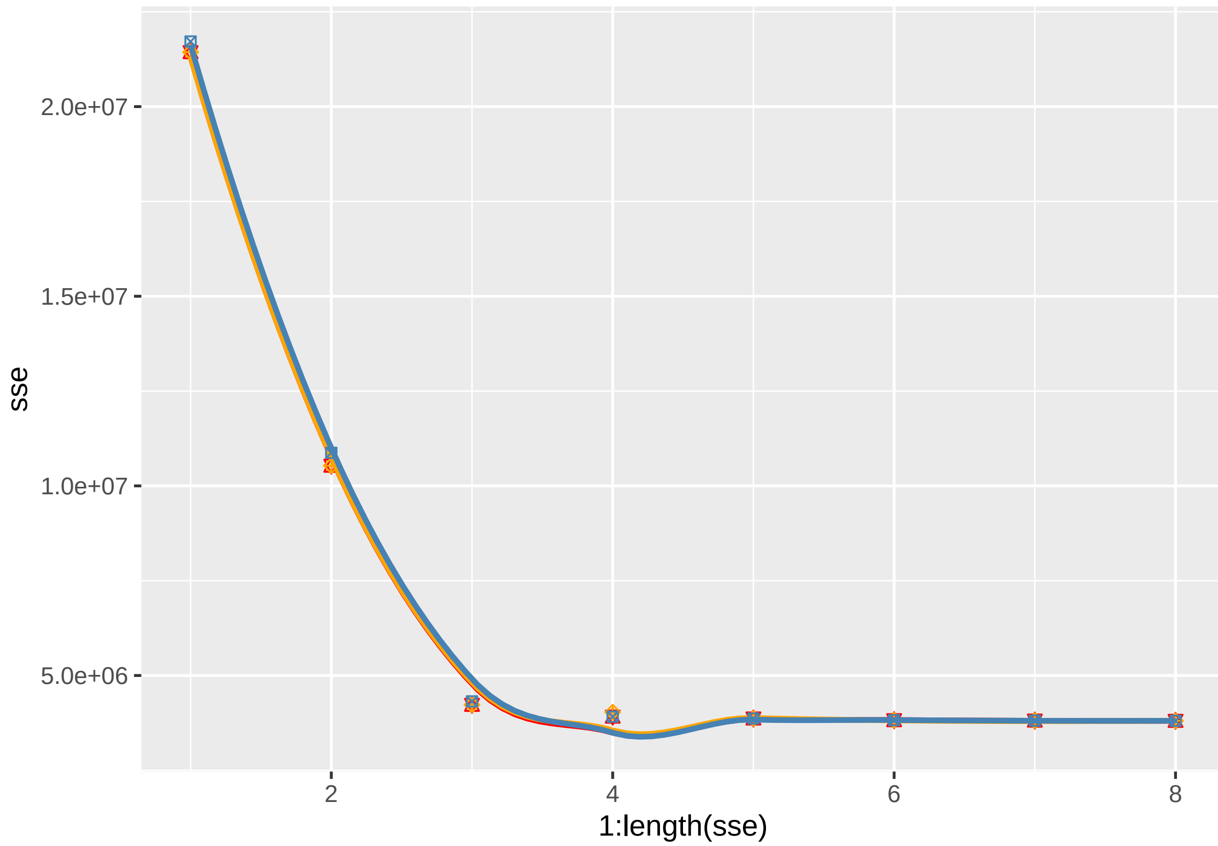
```
## 3  ( 1 ) "*"    "*"   " "    " "   " " " "       " "          "*"
## 4  ( 1 ) "*"    "*"   " "    "*"   " " " "       " "          "*"
## 5  ( 1 ) "*"    "*"   "*"    "*"   " " " "       " "          "*"
## 6  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       " "          "*"
## 7  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"          "*"
## 8  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"          "*"
##          MarriedYes EthnicityAsian EthnicityCaucasian
## 1  ( 1 ) " "        " "            " "
## 2  ( 1 ) " "        " "            " "
## 3  ( 1 ) " "        " "            " "
## 4  ( 1 ) " "        " "            " "
## 5  ( 1 ) " "        " "            " "
## 6  ( 1 ) " "        " "            " "
## 7  ( 1 ) " "        " "            " "
## 8  ( 1 ) " "        "*"            " "
```

```r
backward_sse <- summary(backward_credit)$rss
backward_sse
```

```
## [1] 21715657 10870832  4316997  3915058  3866091  3821620  3810759  3804746
```

```r
ggplot() +
  geom_point(data = data.frame(sse), aes(x=1:length(sse),y=sse), shape=11, color="red")+
  geom_smooth(data = data.frame(sse), aes(x=1:length(sse),y=sse),se=FALSE,color="red")+
  geom_point(data = data.frame(forward_sse), aes(x=1:length(sse),y=forward_sse),shape=9, color="orange")
  geom_smooth(data = data.frame(forward_sse), aes(x=1:length(sse),y=forward_sse),se=FALSE,color="orange
  geom_point(data = data.frame(backward_sse), aes(x=1:length(sse),y=backward_sse),shape=7, color="steel
  geom_smooth(data = data.frame(backward_sse), aes(x=1:length(sse),y=backward_sse),se=FALSE,color="stee
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Part2:

```
best_subsect_Cp <- summary(best_subset_credit)$cp
best_subsect_bic <- summary(best_subset_credit)$bic
best_subsect_Cp;
```

```
## [1] 1800.308406  685.196514   41.133867   11.148910    8.131573    5.574883
## [7]    6.462042    7.845931
```

```
best_subsect_bic;
```

```
## [1]  -535.9468  -814.1798 -1173.3585 -1198.0527 -1197.0957 -1195.7321 -1190.8790
## [8] -1185.5192
```

```
optimal_best_subsect_Cp <- which.min(best_subsect_Cp)
optimal_best_subsect_bic <- which.min(best_subsect_bic)
print(summary(best_subset_credit)$which[optimal_best_subsect_Cp, ])
```

```
##        (Intercept)             Income              Limit             Rating
##               TRUE               TRUE               TRUE               TRUE
##              Cards                Age          Education       GenderFemale
##               TRUE               TRUE              FALSE              FALSE
##         StudentYes         MarriedYes     EthnicityAsian EthnicityCaucasian
##               TRUE              FALSE              FALSE              FALSE
```

```
print(summary(best_subset_credit)$which[optimal_best_subsect_bic, ])
```

```
##        (Intercept)             Income              Limit             Rating
##               TRUE               TRUE               TRUE              FALSE
```

```
##               Cards                 Age           Education        GenderFemale
##                TRUE               FALSE               FALSE               FALSE
##          StudentYes          MarriedYes      EthnicityAsian  EthnicityCaucasian
##                TRUE               FALSE               FALSE               FALSE
```

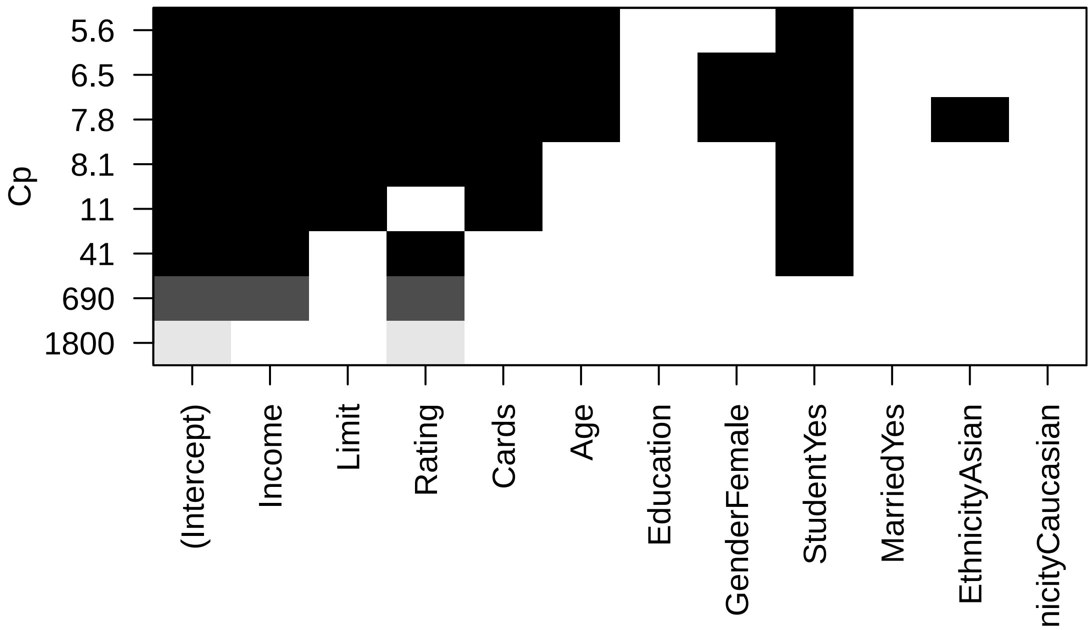
```
optimal_best_subsect_Cp;
```

```
## [1] 6
```

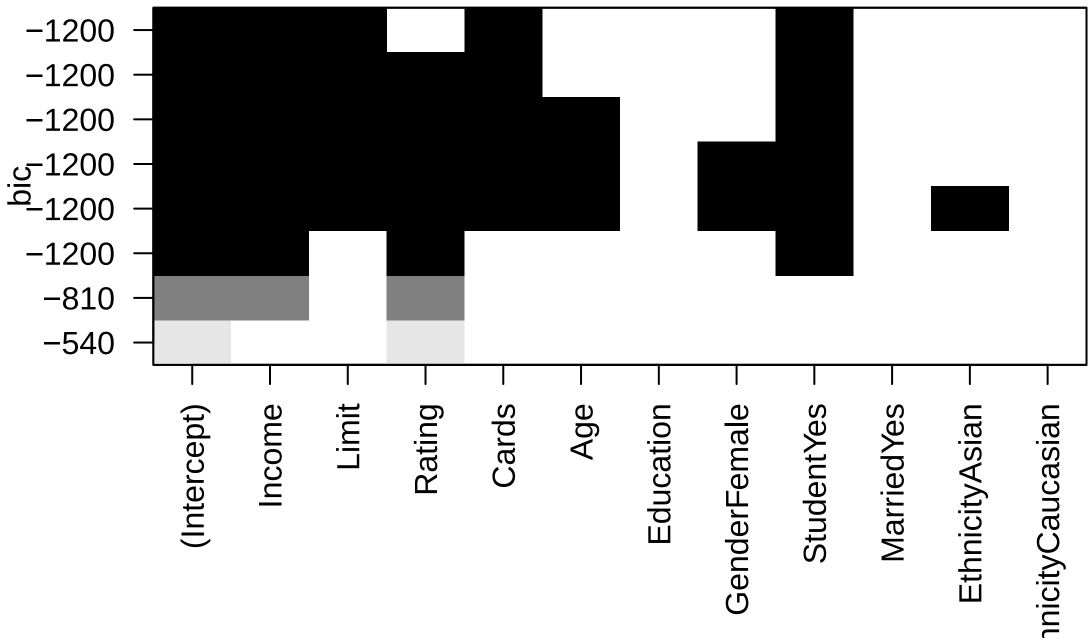
```
optimal_best_subsect_bic
```

```
## [1] 4
```

```
plot(best_subset_credit, scale = "Cp")
```




```
plot(best_subset_credit,scale = "bic")
```




```r
forward_Cp <- summary(forward_credit)$cp
forward_bic <- summary(forward_credit)$bic
forward_Cp;
```

```
## [1] 1800.308406  685.196514   41.133867   23.182500    8.131573    5.574883
## [7]    6.462042    7.845931
```

```r
forward_bic;
```

```
## [1]  -535.9468  -814.1798 -1173.3585 -1186.2300 -1197.0957 -1195.7321 -1190.8790
## [8] -1185.5192
```

```r
optimal_forward_Cp <- which.min(forward_Cp)
optimal_forward_bic <- which.min(forward_bic)
print(summary(forward_credit)$which[optimal_forward_Cp, ])
```

```
##       (Intercept)            Income             Limit            Rating
##              TRUE              TRUE              TRUE              TRUE
##             Cards               Age         Education      GenderFemale
##              TRUE              TRUE             FALSE             FALSE
##        StudentYes        MarriedYes    EthnicityAsian EthnicityCaucasian
##              TRUE             FALSE             FALSE             FALSE
```

```r
print(summary(forward_credit)$which[optimal_forward_bic, ])
```

```
##       (Intercept)            Income             Limit            Rating
##              TRUE              TRUE              TRUE              TRUE
##             Cards               Age         Education      GenderFemale
##              TRUE             FALSE             FALSE             FALSE
##        StudentYes        MarriedYes    EthnicityAsian EthnicityCaucasian
##              TRUE             FALSE             FALSE             FALSE
```

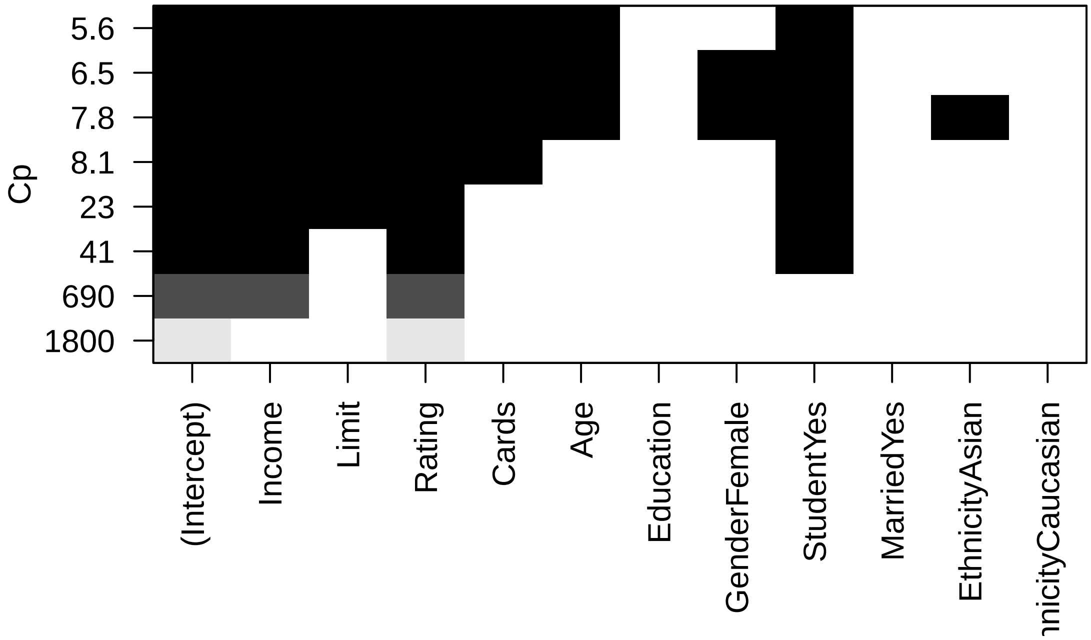
```r
optimal_forward_Cp;
```

```
## [1] 6
```

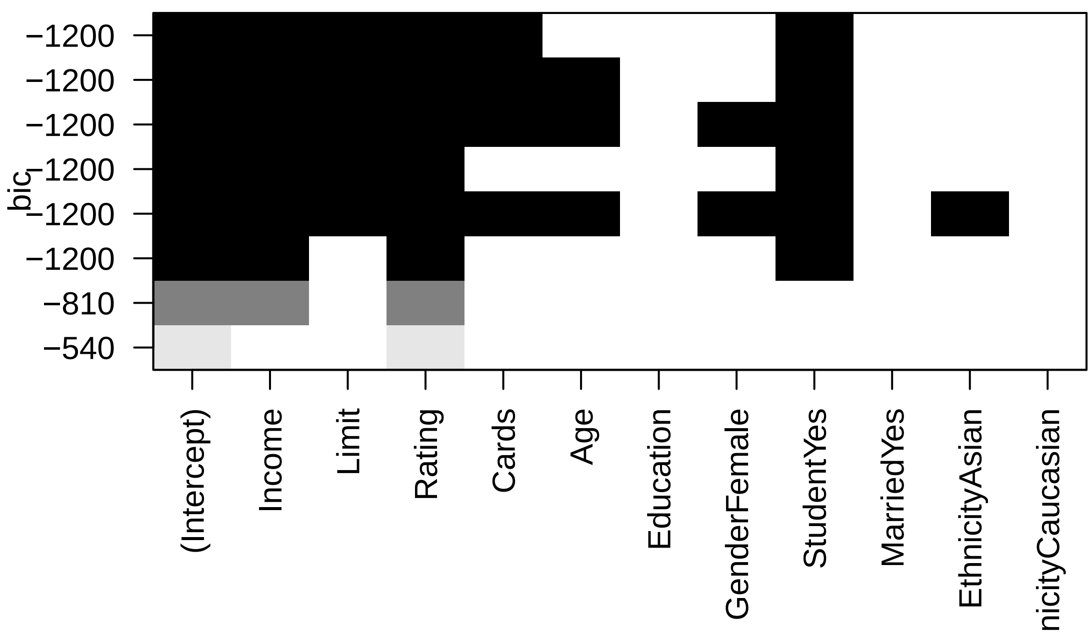
```r
optimal_forward_bic
```

```
## [1] 5
```

```r
plot(forward_credit, scale = "Cp")
```




```r
plot(forward_credit,scale = "bic")
```




```r
backward_Cp <- summary(backward_credit)$cp
backward_bic <- summary(backward_credit)$bic
backward_Cp;
```

```
## [1] 1829.052845  719.858831   50.332736   11.148910    8.131573    5.574883
## [7]    6.462042    7.845931
```

```r
backward_bic;
```

```
## [1]  -530.7458  -801.5344 -1164.9522 -1198.0527 -1197.0957 -1195.7321 -1190.8790
## [8] -1185.5192
```

```r
optimal_backward_Cp <- which.min(backward_Cp)
optimal_backward_bic <- which.min(backward_bic)
print(summary(backward_credit)$which[optimal_backward_Cp, ])
```

```
##        (Intercept)             Income              Limit             Rating
##               TRUE               TRUE               TRUE               TRUE
##              Cards                Age          Education       GenderFemale
##               TRUE               TRUE              FALSE              FALSE
##         StudentYes         MarriedYes     EthnicityAsian EthnicityCaucasian
##               TRUE              FALSE              FALSE              FALSE
```

```r
print(summary(backward_credit)$which[optimal_backward_bic, ])
```

```
##        (Intercept)             Income              Limit             Rating
##               TRUE               TRUE               TRUE              FALSE
##              Cards                Age          Education       GenderFemale
##               TRUE              FALSE              FALSE              FALSE
##         StudentYes         MarriedYes     EthnicityAsian EthnicityCaucasian
##               TRUE              FALSE              FALSE              FALSE
```

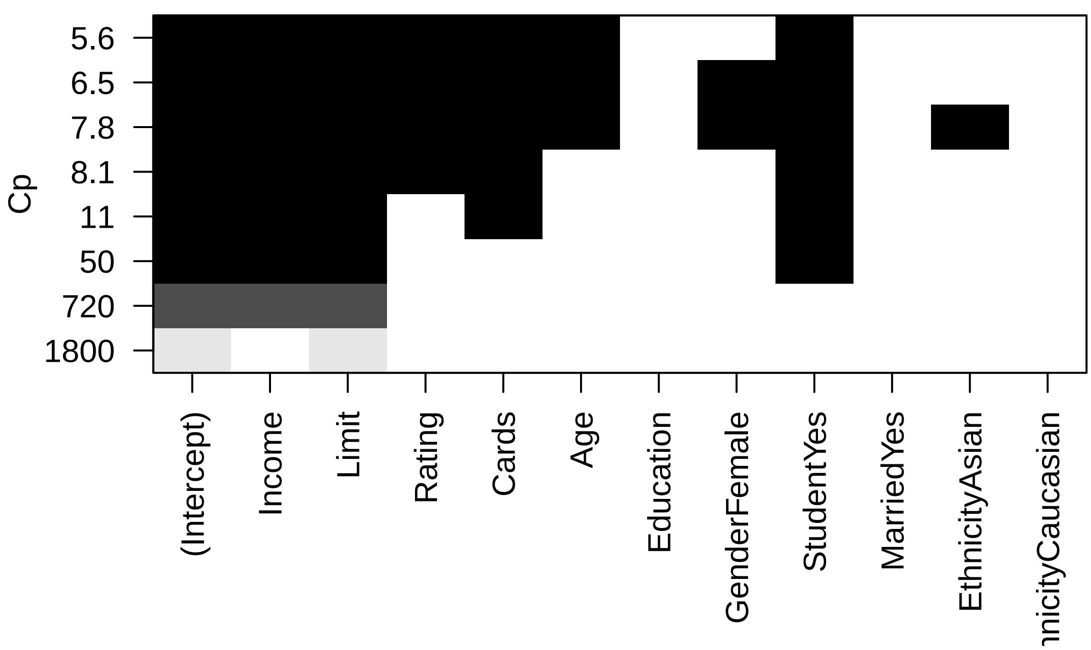
```r
optimal_backward_Cp;
```

```
## [1] 6
```

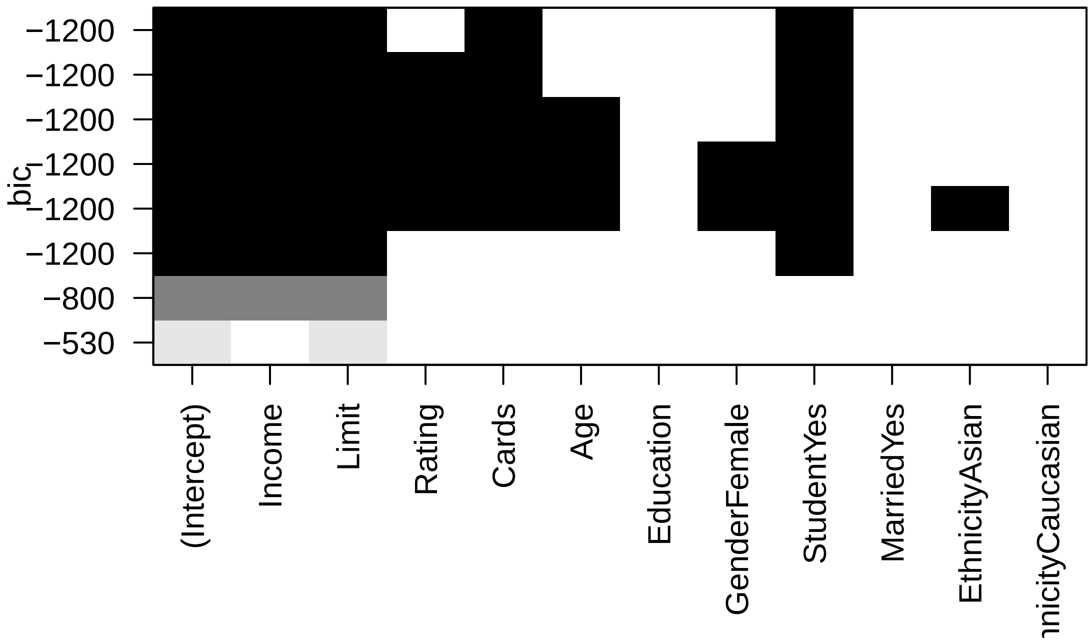
```r
optimal_backward_bic
```

```
## [1] 4
```

```r
plot(backward_credit, scale = "Cp")
```




```r
plot(backward_credit,scale = "bic")
```



Now we have 3 subset select methods, for best subset procedure: optimal model is Balance~Income+Limit+Rating+Cards+Age+ by using Cp and we have 6 predictors; optimal model is Balance~Income+Limit+Cards+StudentYes, by using BIC and we have 4 predictors.

For forward select: optimal model is Balance~Income+Limit+Rating+Cards+Age+StudentYes, by using Cp and we have 6 predictors; optimal model is Balance~Income+Limit+Rating+Cards+StudentYes, by using BIC and we have 5 predictors.

For backward select: optimal model is Balance~Income+Limit+Rating+Cards+Age+StudentYes, by using Cp and we have 6 predictors; optimal model is Balance~Income+Limit+Cards+StudentYes, by using BIC and we have 4 predictors.